Je parle à mon moteur de recherche et il me répond

Stéphane Bortzmeyer
<stephane+blog@bortzmeyer.org>

Première rédaction de cet article le 2 mai 2007

`http://www.bortzmeyer.org/je-parle-a-mon-moteur-de-recherche.html`

Lire les journaux de son serveur HTTP est une source d'amusement sans fin. En effet, beaucoup d'utilisateurs surestiment nettement leur moteur de recherche favori et lui parlent en langue naturelle, entrainant parfois des résultats suprenants.

Malheureusement, l'apprentissage de l'usage d'Internet se limite en général à l'apprentissage des outils (≪ clique ici ≫), pas à leur utilisation critique. L'utilisateur naïf est donc voué à faire des erreurs, qui distraient l'informaticien occupé à utiliser `tail` ou l'excellent xlogmaster `<http://www.gnu.org/software/xlogmaster/>` sur les journaux d'Apache. La plupart du temps, l'utilisateur confond le moteur de recherche avec un robot d'un film de science-fiction des années 50, à qui on peut parler comme s'il était humain.

Voici donc quelques requêtes particulièrement intéressantes vues sur mon blog `<http://www.bortzmeyer.org/>`, avec les réflexions qu'elles m'inspirent. L'orthographe originale a été strictement respectée.

`getaddrinfo exemple utilisation` (a mené en `<http://www.bortzmeyer.org/mesurer-temps-executi html>`). La page en question parle bien de `getaddrinfo` mais ne contient pas d'exemple, le mot ≪ exemple ≫ apparaissait pour autre chose. À l'inverse, une page avec un exemple d'utilisation de `getaddrinfo` peut ne pas comporter du tout le mot ≪ exemple ≫. Cette erreur est sans doute la plus courante : croire que le moteur de recherche manipule des concepts alors qu'il ne connait que des chaînes de caractères. Il ne sait pas que, par exemple, ≪ vin ≫ est synonyme de ≪ pinard ≫. Le cas est d'autant plus troublant ici que l'utilisation de `getaddrinfo` indique que l'utilisateur était un informaticien... Il aurait plutôt dû utiliser un moteur de recherche spécialisé dans le code source comme Code Search `<http://www.google.com/codesearch>` ou Krugle `<http://www.krugle.com/>`. Encore mieux, après la publication de cet article, c'est celui-ci qu'a trouvé un autre étudiant cherchant `getaddrinfo exemple`. Et celui-ci venait d'une grande école d'ingénieurs de Grenoble, ce qui indique un manque de formation à l'usage des outils du réseau dans ces établissements...

`PostgreSQL comment cela marche` (a mené en `<http://www.bortzmeyer.org/postgresql-unicode.html>`). Encore pire puisqu'il y a peu de chance que le moteur de recherche aie compris le ≪ comment ça marche ≫. L'utilisateur aurait plutôt dû partir de l'article de Wikipédia, qui indique plusieurs tutoriels.

empécher IE de faire une requête DNS (a mené en <http://www.bortzmeyer.org/fermer-les-r html>). Là encore, la syntaxe de la phrase a complètement échappé au moteur de recherche qui a simplement trouvé une page où tous ces mots apparaissaient. Mettre la même phrase avec les mots dans le désordre donne d'ailleurs presque les mêmes résultats. Plus drôle, revendeur d'objets polonais a mené en <http://www.bortzmeyer.org/3730.html> car l'analyse du RFC parlait en effet des polonais, d'objets et de revendeurs (sans que ces termes soient liés).

ecrire un roman avec microsoft word (a mené en <http://www.bortzmeyer.org/afterword.html>). L'utilisateur a dû être très déçu de trouver un texte critiquant vigoureusement Word (cet article reçoit énormément de visites d'utilisateurs de Word, ce qui indique bien le manque de documentation sur cet outil prétendûment simple d'utilisation). Son problème est que la syntaxe très limitée du langage de requêtes des moteurs de recherche ne permet pas d'exprimer des concepts comme le fait que la page soit orientée vers les utilisateurs de Word. Le moteur de recherche a juste vu le mot ≪ word ≫. Même si les moteurs de recherche étaient plus perfectionnés, les pages Web ne sont typiquement pas structurées et il serait très difficile d'en déduire automatiquement si une page est un mode d'emploi ou un article polémique.

blog sur les strings (a mené en <http://www.bortzmeyer.org/3454.html>). Sans doute encore une grande déception pour l'utilisateur, qui n'a pas pensé que le mot "*string*" était répandu en franglais informatique, pas seulement en lingerie.

pourquoi tant de gens utilisent-ils internet? (a mené en <http://www.bortzmeyer.org/identificateur-vs-moteur-de-recherche.html>). Excellente question mais certainement trop philosophique pour un moteur de recherche. Le lycéen paresseux qui l'a tapée dans Google aura peu de chance d'avoir une réponse toute prête pour sa dissertation.

expresso capucino difference (a mené en <http://www.bortzmeyer.org/data-formats.html>). Outre la faute d'orthographe à cappuccino, l'utilisateur est tombé victime d'un problème courant, les exemples. Les informaticiens aiment utiliser des noms pittoresques et les exemples dans mes articles me valent des visites inattendues par exemple un amateur de Ragnarok qui tombe sur <http://www.bortzmeyer.org/xen.html> ou bien un client de Véolia qui arrive en <http://www.bortzmeyer.org/2672.html>. J'ai aussi vu nom et prenom de la physicienne polonaise arriver en <http://www.bortzmeyer.org/postgresql-unicode.html>.

images e t textes blog sur la mort (a mené en <http://www.bortzmeyer.org/afterword.html>). Un sujet sinistre et un résultat inattendu (mais où le mot mort figurait bien).

mon moteur de recherche (a mené à cet article), est une requête apparue depuis la publication de la première version de cet article.

L'excellente étude ≪ Usages de l'Internet par les étudiants burkinabé <http://www.tic.ird.fr/article.php?id_article=252> ≫ montre bien le phénomène et, comme le montrent les requêtes plus haut (toutes venues de France), cela n'a rien de spécifique au Burkina-Faso. Ainsi, dans l'enquête, une documentaliste se plaint que, pour les étudiants, ≪ Google est le point d'entrée quasi général. Mais si certains entrent des mots-clés, beaucoup indiquent la référence complète d'un ouvrage et s'étonnent d'avoir en réponse, une liste de librairies en ligne. De même, beaucoup posent leur question par **une longue phrase en langage naturel dont les mots peu significatifs amène Google à retourner des réponses non pertinentes** ≫.

Finissons sur une note optimiste. Voici quelques requêtes qui ont bien marché, peut-être par pure chance, mais tant mieux pour leur auteur.

— Convertisseur XML vers csv (a mené en <http://www.bortzmeyer.org/xml-to-csv.html>).
— Host Identity Protocol (a mené en <http://www.bortzmeyer.org/4423.html>).
— dhclient bail (a mené en <http://www.bortzmeyer.org/2131.html>).
— le maitre de garamond (a mené en <http://www.bortzmeyer.org/garamond.html>).

Cette étude a été faite en examinant le journal d'un serveur Apache configuré ainsi :

```
LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}i\"" combined
CustomLog /web/logs/bortzmeyer/access.log combined
```

Le journal a été analysé par ce programme (en ligne sur http://www.bortzmeyer.org/files/SearchEngineQueries.py).

D'autres auteurs de sites Web ont déjà fait des constatations similaires comme Michel Fingerhut <http://mmdl.free.fr/blog-m/?p=330> qui publie ces amusants exercices de style <http://mmdl.free.fr/blog-m/?p=406> ou comme Romain Vimont <http://blog.rom1v.com/2011/06/extraire-les-recherches-google-des-logs-apache/>. La naïveté des utilisateurs à l'égard des moteurs de recherche a été très bien montrée dans le petit film des Deux minutes du peuple, <http://www.youtube.com/watch?v=82wnezAMKu0>.

Et, bien sûr, la meilleure illustration des limites des moteurs de recherche, se trouve dans le film de Spielberg, A.I. où le héros demande au logiciel « Dr. Know » où trouver la Fée Bleue et où le moteur de recherche, ignorant du contexte, croit qu'il s'agit d'une fleur <http://www.candcgardens.com/Flower%20Pages/blue_fairy.htm>...